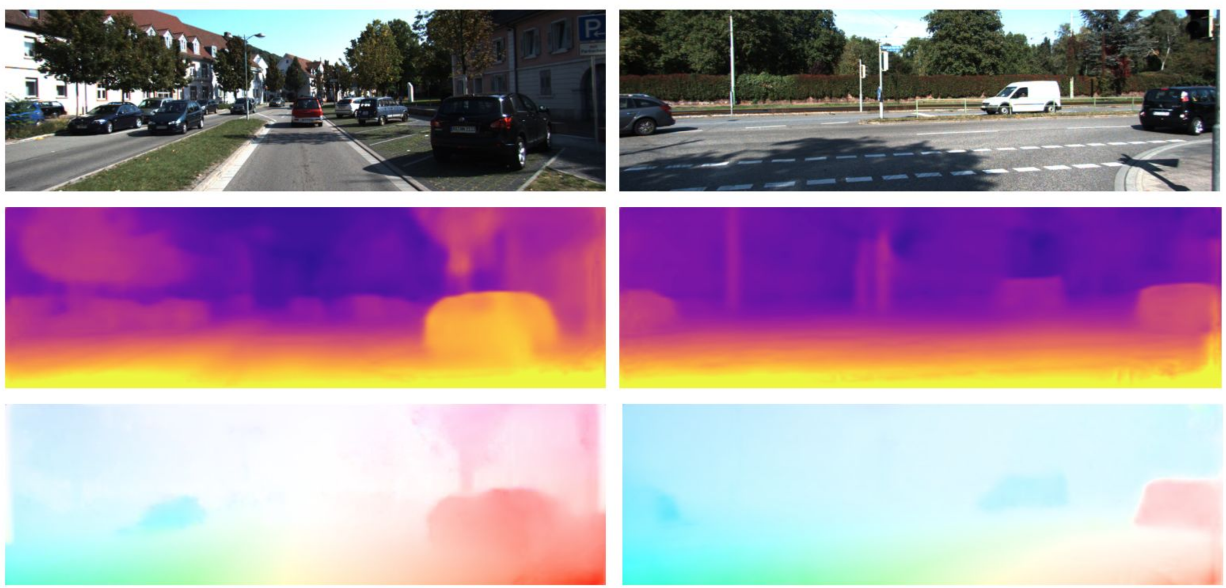
GeoNet: Unsupervised Learning of Dense Depth, Optical Flow and Camera Pose

Presented by **Owen Jow**
Original work by **Zhichao Yin** and **Jianping Shi**

Motivation and Problem Description

- Densely estimate **depth** / **camera motion** / **optical flow** from monocular video
- Useful for self-driving cars and robots (e.g.)



input image

depth map

optical flow

Prior Work

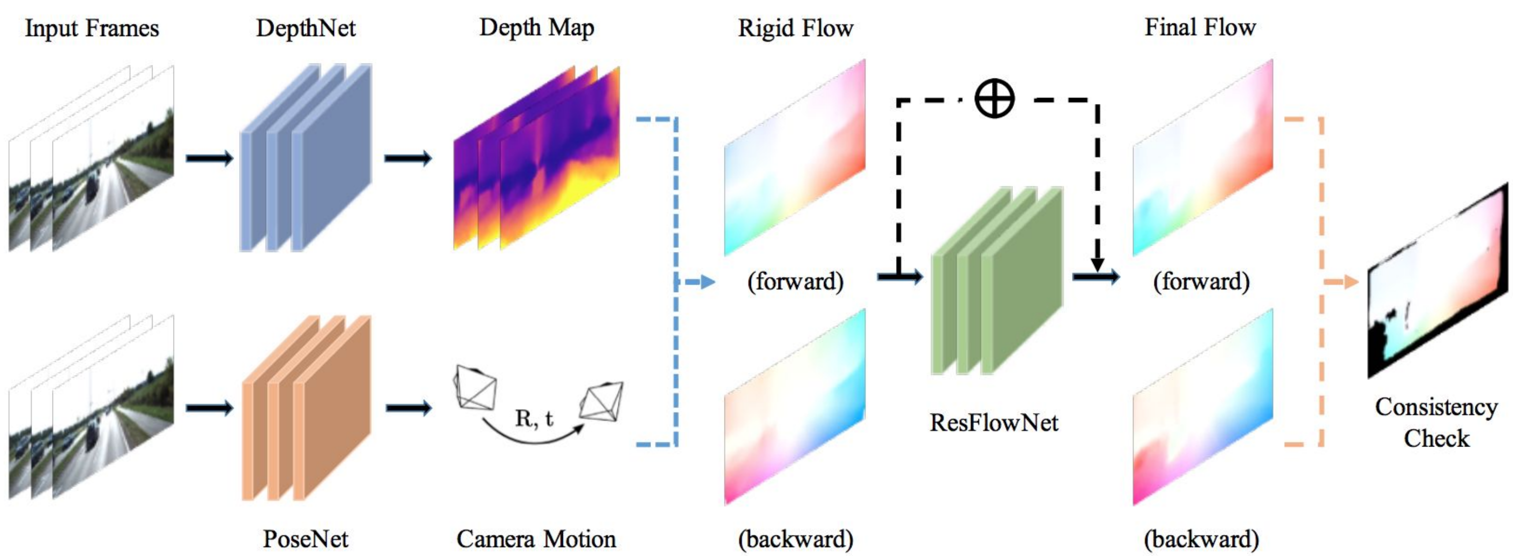
- **Traditional methods**
  - SfM for depth and camera motion, Lucas-Kanade (e.g.) for optical flow
  - **But** reliant on texture and photo-consistency for correspondence, difficult/slow optimization...
- **Deep supervised learning**
  - Train network using input/output data → successes (like FlowNet 2.0) for all three problems
  - Network learns to identify and exploit cues at both low and high levels
  - **But** reliant on supervision (expensive to collect), conventionally performs task in isolation...
- **Deep unsupervised learning**
  - Use image reconstruction objective in lieu of direct supervision
  - **But** prior systems don't exploit geometric consistency, manage occlusion/dynamic objects...

Method Overview

- 1. Estimate **Static** Scene Geometry
  - Estimate depth maps and camera motion for and between frames
  - Combine to get rigid flow field, then can warp source view to target (+minimize diff with real)

$$f^{rig}_{t\to s}(p_t) = KT_{t\to s}D_t(p_t)K^{-1}p_t - p_t$$

- 2. Refine Motion Based on **Dynamic** Scene Geometry
  - Estimate **additive refinement** to rigid flow field, then can use to synthesize target view again



Method Details and Analysis

- Minimize sum of loss terms over multiple scales and source/target pairs
  - **Rigid/full flow warping**: perceptual and L1 loss between synthesized and actual view
  - **Depth/full flow smoothness**: minimize gradients in low-frequency image regions
  - **Geometric consistency**: forward flow + backward flow should return to same pixel
  - Adaptive consistency weighting: only impose **full flow warping loss** and **geometric consistency loss** at a pixel if the `(s⟶t flow + t⟶s flow)` is below some threshold

Method Details and Analysis

- Minimize sum of loss terms over multiple scales and source/target pairs
  - **Rigid**/full **flow warping**: perceptual and L1 loss between synthesized and actual view
  - **Depth**/full flow **smoothness**: minimize gradients in low-frequency image regions
  - **Geometric consistency**: forward flow + backward flow should return to same pixel
  - Adaptive consistency weighting: only impose **full flow warping loss** and **geometric consistency loss** at a pixel if the `(s⟶t flow + t⟶s flow)` is below some threshold
- Forward-backward check eliminates the need to predict flows involving occluded, texture-ambiguous, or otherwise photo-inconsistent regions
  - Interpolate the values from the rest of the flow field; don't enforce reconstruction loss

Method Details and Analysis

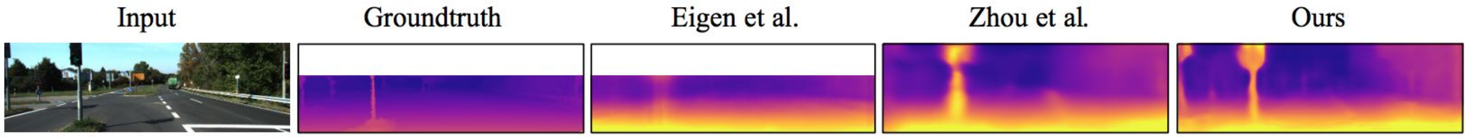
- Minimize sum of loss terms over multiple scales and source/target pairs
    - **Rigid/full flow warping**: perceptual and L1 loss between synthesized and actual view
    - **Depth/full flow smoothness**: minimize gradients in low-frequency image regions
    - **Geometric consistency**: forward flow + backward flow should return to same pixel
    - Adaptive consistency weighting: only impose **full flow warping loss** and **geometric consistency loss** at a pixel if the (`s→t flow + t→s flow`) is below some threshold
- Forward-backward check eliminates the need to predict flows involving occluded, texture-ambiguous, or otherwise photo-inconsistent regions
    - Interpolate the values from the rest of the flow field; don't enforce reconstruction loss
- Rigid/static reconstructor provides good starting point for flow field
    - [Stage 1] rigidly-constrained flow field → [Stage 2] unconstrained flow field
    - Residual flow network can refine incorrect predictions and deal with dynamic objects

Method Details and Analysis

- Minimize sum of loss terms over multiple scales and source/target pairs
  - **Rigid**/full **flow warping**: perceptual and L1 loss between synthesized and actual view
  - **Depth**/full flow **smoothness**: minimize gradients in low-frequency image regions
  - **Geometric consistency**: forward flow + backward flow should return to same pixel
  - Adaptive consistency weighting: only impose **full flow warping loss** and **geometric consistency loss** at a pixel if the `(s→t flow + t→s flow)` is below some threshold
- Forward-backward check eliminates the need to predict flows involving occluded, texture-ambiguous, or otherwise photo-inconsistent regions
  - Interpolate the values from the rest of the flow field; don't enforce reconstruction loss
- Rigid/static reconstructor provides good starting point for flow field
  - [Stage 1] rigidly-constrained flow field → [Stage 2] unconstrained flow field
  - Residual flow network can refine incorrect predictions and deal with dynamic objects
- Preferably train in two stages (first DepthNet/PoseNet, then ResFlowNet)

Experiments

- Evaluate on predefined data splits (with GT) for KITTI driving dataset
- Better than previous unsupervised, comparable to previous supervised
- **Depth estimation**
  - Worse than supervised method (Godard et al.) at resolving dataset differences



Method	Supervised	Dataset	Abs Rel	Sq Rel	RMSE
Eigen *et al.* [9] Coarse	Depth	K	0.214	1.605	6.563
Eigen *et al.* [9] Fine	Depth	K	0.203	1.548	6.307
Liu *et al.* [28]	Depth	K	0.202	1.614	6.523
Godard *et al.* [15]	Pose	K	**0.148**	1.344	5.927
Zhou *et al.* [56]	No	K	0.208	1.768	6.856
Zhou *et al.* [56] updated[2]	No	K	0.183	1.595	6.709
Ours VGG	No	K	0.164	1.303	6.090
Ours ResNet	No	K	0.155	**1.296**	**5.857**
Garg *et al.* [14] cap 50m	Pose	K	0.169	1.080	5.104
Ours VGG cap 50m	No	K	0.157	0.990	4.600
Ours ResNet cap 50m	No	K	**0.147**	**0.936**	**4.348**
Godard *et al.* [15]	Pose	CS + K	**0.124**	**1.076**	**5.311**
Zhou *et al.* [56]	No	CS + K	0.198	1.836	6.565
Ours ResNet	No	CS + K	0.153	1.328	5.737

Experiments

- Evaluate on predefined data splits (with GT) for KITTI driving dataset
- Better than previous unsupervised, comparable to previous supervised
- **Optical flow estimation**
  - Endpoint error
  - Validate the use of residual flow and adaptive geometric consistency
  - GeoNet better at fixing **small** rigid flow errors; pixel intensity contrast loss is inherently local

Method	Dataset	Noc	All
EpicFlow [38]	-	4.45	9.57
FlowNetS [8]	C+S	8.12	14.19
FlowNet2 [18]	C+T	4.93	10.06
DSTFlow [37]	K	6.96	16.79
Our DirFlowNetS (no GC)	K	6.80	12.86
Our DirFlowNetS	K	**6.77**	12.21
Our Naive GeoNet	K	8.57	17.18
Our GeoNet	K	8.05	**10.81**

Experiments

- Evaluate on predefined data splits (with GT) for KITTI driving dataset
- Better than previous unsupervised, comparable to previous supervised

- **Camera motion estimation**
    - Absolute [camera frame] trajectory error
    - Compare against ORB-SLAM and unsupervised SfM method (Zhou et al.)

Method	Seq.09	Seq.10
ORB-SLAM (full)	0.014 ± 0.008	0.012 ± 0.011
ORB-SLAM (short)	0.064 ± 0.141	0.064 ± 0.130
Zhou *et al.* [56]	0.021 ± 0.017	0.020 ± 0.015
Zhou *et al.* [56] updated	0.016 ± 0.009	0.013 ± 0.009
Our GeoNet	$\mathbf{0.012 \pm 0.007}$	$\mathbf{0.012 \pm 0.009}$

Future Work and Discussion

- Future work:
  - Introduce semantic information
  - Avoid gradient locality of warping loss
  - Leverage temporal consistency to a greater degree, e.g. depth prediction is single view
- **Can exploit geometric relationships between depth, optical flow, and camera motion to train jointly in an unsupervised, end-to-end fashion**